Supplementary Material for HyperTransformer: A Textural and Spectral Feature Fusion Transformer for Pansharpening

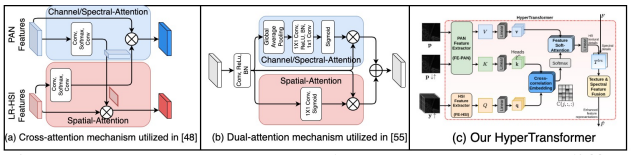

Figure 7: How our proposed *HyperTransformer* differs from the attention mechanisms utilized in previous pansharpening works.

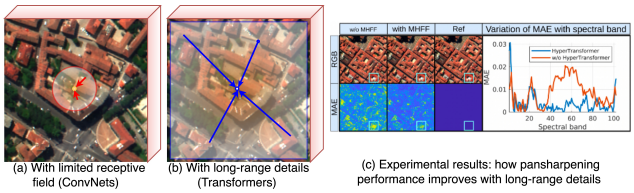

Figure 9: Necessity of long-range details for pansharpening.

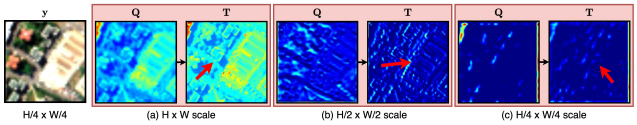

Figure 8: For a given query feature **Q**, we visualize corresponding spectrally similar and texturally advanced feature map **T** processed from our *HyperTransformer* at multiple spatial scales.

How our proposed HyperTransformer differs from previous pansharpening methods that utilize attention? In previous pansharpening works [48, 55], attention (channel and spatial) mechanisms are used to *re-weight* the PAN and LR-HSI features from ConvNets along the channel and spatial dimensions as shown in Fig. 7 (a) and (b) without having an explicit consideration of special properties of PAN and LR-HSI features. Different from these previous methods, the proposed HyperTransformer is specifically designed to cater to the pansharpening problem by taking into consideration spatial and spectral properties of PAN and LR-HSI. Instead of simply re-weighting the feature maps, our HyperTransformer first computes the cross-correlation between the feature representations of PAN$\downarrow\uparrow$ and LR-HSI. Then multi-head feature soft-attention (MHFA) is utilized to identify texturally advanced and spectrally similar feature representations from PAN that will be further mixed with spectral features from the backbone network. Hence, the proposed HyperTransformer re-defines queries, keys, and values in standard attention mechanisms as LR-HSI, PAN$\downarrow\uparrow$, and PAN features, respectively that not only deliver better intuitive understanding to the pansharpening problem under the context of attention but also result in better pansharpening performance. HyperTransformer outperforms many previous classical [1, 2, 3, 16, 18, 23, 25, 33, 40, 41, 52], ConvNet-based [5, 12, 22, 44, 46], and attention-based [55] methods in terms of CC, SAM, RSNR, ERGAS, and PSNR on three datasets.

Visualization of *in* and *out* feature maps from our HyperTransformer. As shown in Fig. 8, at each spatial scale, HyperTransformer adds missing texture details to LR-HSI features (queries - **Q**) while maintaining their spectral characteristics (i.e., the cross-correlation).

Why are long-range details necessary for pansharpening? We explain our intuition of why pansharpening should benefit from long-range details in Fig. 9. As shown in Fig. 9, when the pansharpening network has a larger receptive field (i.e., it can capture long-range details), it can enhance the texture and spectral details of a given pixel not only by looking at adjacent pixels but also from the pixels far away. As shown in 9 - (c) (in paper Fig. 4), we can see a significant reduction in MAE across the spectral bands when we add our HyperTransformer to the main pansharpening network, which empirically shows that pansharpening indeed benefits from long-rage details. Furthermore, it has been shown in the literature that not only segmentation, detection, and classification tasks benefit from long-rage details but also restoration, fusion, and super-resolution [1]. Pansharpening consists of both super-resolution and fusion

tasks and as a result it should also benefit from long-range details.

References

[1] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Comput. Surv.*, dec 2021.